Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 141 (2014) 124 – 128

Procedia
Social and Behavioral Sciences

WCLTA 2013

Using Corpus Linguistics in the Development of Writing

Blanka Frydrychova Klimova*

University of Hradec Kralove, Faculty of Informatics and Management, Rokitanskeho 62, Hradec Kralove, 500 03, Czech Republic

Abstract

Corpus linguistics has its adequate justification in the teaching of writing skills since it can reveal patterns of authentic language use through analyses of actual usage. The aim of this paper is to list the most common corpora and explore useful, inexpensive and user-friendly software programmes, such as the WordSmith Tools or TextSTAT, exploited in the development of writing skills. In addition, the author provides several practical examples on how corpus linguistics can be applied to the development of writing skills. For instance, the concordance enables to see any word or phrase in context so that one can see what sort of company it keeps. Thus, students can, for example, see the differences between the words they often confuse (e.g., *excited* vs. *exciting*).

Keywords: Corpus, concordancing, writing, language use;

1. Main text

Corpus linguistics is the study of language based on examples of *real life* language use stored in corpora (or corpuses) - computerized databases created for linguistic research. It is the study of language through computational analyses of large collections of written texts and recordings of speech.

Based on the above definition of the corpus, corpus linguistics is the study of language by means of naturally occurring language samples; analyses are usually carried out with specialised software programmes on a computer. Corpus linguistics is thus a method to obtain and analyse data quantitatively and qualitatively rather than a theory of language. The corpus-linguistic approach can be used to describe language features and to test hypotheses formulated in various linguistic frameworks. To name but a few examples, corpora recording different stages of

* Corresponding Author: Blanka Frydrychova Klimova. Tel.: +420 493-332-318
E-mail address:blankaklimova@uhk.zu

doi:10.1016/j.sbspro.2014.05.023

learner language (beginners, intermediate, and advanced learners) can provide information for foreign language acquisition research; by means of historical corpora it is possible to track the development of specific features in the history of English like the emergence of the modal verbs *gonna* and *wanna*; or sociolinguistic markers of specific age groups such as the use of *like* as a discourse marker can be investigated for purposes of sociolinguistic or discourse-analytical research.

The great advantage of the corpus-linguistic method is that language researchers do not have to rely on their own or other native speakers' intuition or even on made-up examples. Rather, they can draw on a large amount of authentic, naturally occurring language data produced by a variety of speakers or writers in order to confirm or refute their own hypotheses about specific language features on the basis of an empirical foundation.

There exist different types of corpora. Among these the best-known and influential types are as follows:

General corpora, such as the *British National Corpus* (BNC) or the *Bank of English* (BoE), contain a large variety of both written and spoken language, as well as different text types, by speakers of different ages, from different regions and from different social classes.

Synchronic corpora, such as *F-LOB* and *Frown*, record language data collected at one specific point in time, e.g. written British and American English of the early 1990s.

Historical corpora, such as *A Representative Corpus of Historical English Registers* (ARCHER) and the *Helsinki Corpus of English Texts*, consist of corpus texts from earlier periods of time. They usually span several decades or centuries, thus providing diachronic coverage of earlier stages of language.

Learner corpora, such as the *International Corpus of Learner English* (ICLE) and the *Cambridge Learner Corpus* (CLC), are collections of data produced by foreign language learners, such as essays or written exams.

Corpora for the study of varieties, such as the *International Corpus of English* (ICE) and the *Freiburg English Dialect Corpus*, represent different regional varieties of a language.

There is also a large variety of *specialized corpora*, e.g. *Michigan Corpus of Academic Spoken English* (*MICASE*),

In addition, corpora are then analysed with the help of a concordancing programme which can list the most common patterns and rules of a particular register. Greenbaum (1991: 87) states that for an analysis of professional texts a language corpus of 20,000-30,000 is sufficient. At present, there are a number of such programmes, for example, the WordSmith Tools (Scott, 2012), which is an integrated suite of programs for looking at how words behave in texts (see Fig. 1).

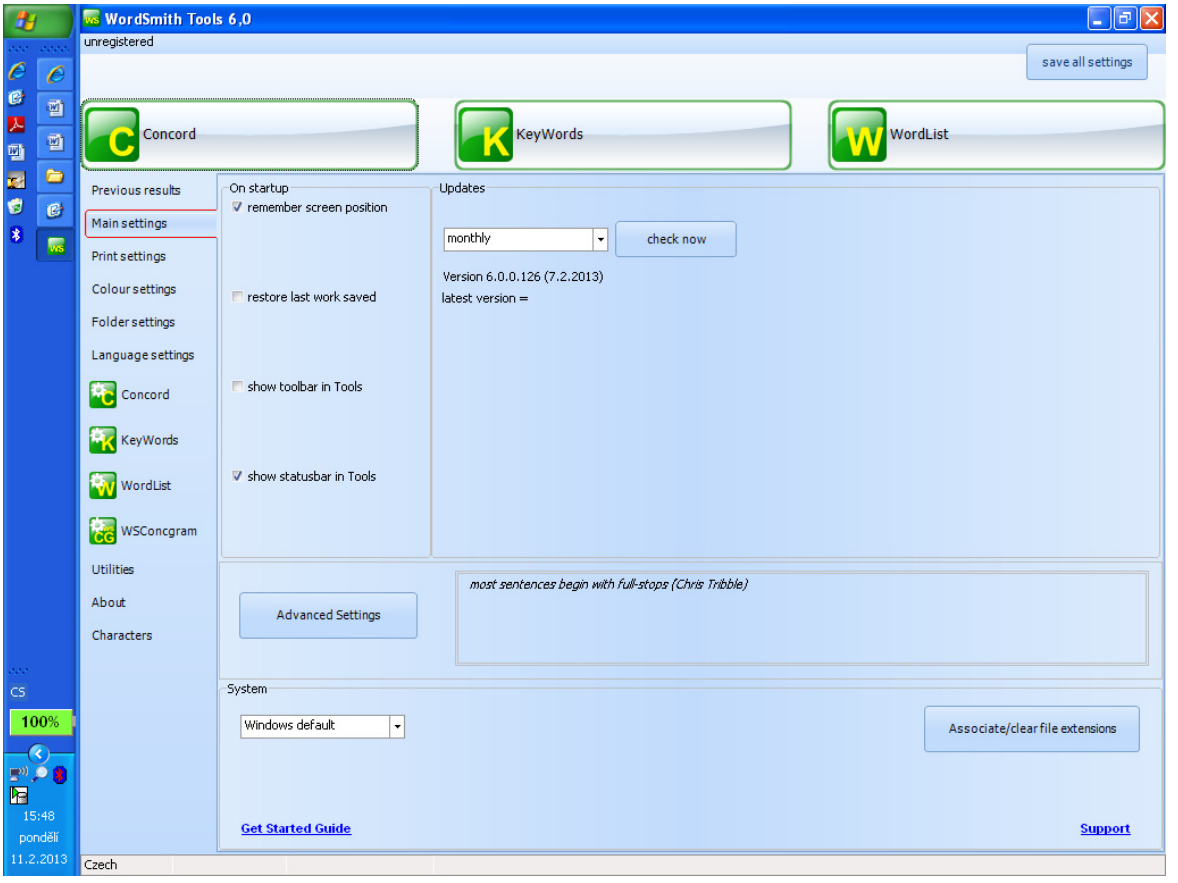

Fig. 1. WordSmith Tools

This programme helps learners find out how words are used in authentic texts and thus how they may be used in one's own texts. The *WordList* tool can provide a list of all the words or word-clusters in a text, set out in alphabetical or frequency order.

The *concordancer, Concord*, gives a chance to see any word or phrase in context so that one can see what sort of company it keeps. This has many advantages for foreign language teachers and learners since they see the differences between words they often confuse (e.g. *excited* vs. *exciting*); they can observe the patterns of various forms; they can see the connotative meanings a word acquires because of its regular association with other words (e.g. the word *commit* carries unfavourable implications because of its regular co-occurrence with words such as *crime, murder,* or *fraud*); they can discover the most appropriate words to use (e.g. whether to use the preposition *in* or *to* with *specialize)*; they can follow the stable lexical patterning in particular disciplines, particularly nominal groups (e.g. *critical discourse analysis*); or they can examine the specific meanings that words take on in particular disciplines (e.g.*, floor, sentence, structure* ,or *wall*). Thus, one must agree with Wu (1992: 32):

> Only when words are in their habitual environments, presented in their most frequent forms and their relational patterns and structures, can they be learnt effectively, interpreted properly and used appropriately.

Finally, with *KeyWords* one can find the key words in a text. Key-words provide a useful way to characterise a text or a genre. The program compares two pre-existing word-lists, which were created using the WordList tool. One of these is assumed to be a large word-list which will act as a reference file. The other is the word-list based on one text which a person wants to study. The aim is to find out which words characterise the text one is most interested

in, which is automatically assumed to be the smaller of the two texts chosen. The larger will provide background data for reference comparison.

However, the WordSmith Tools programme must be paid for. Fortunately, there are other software programmes for the analysis of corpora, such as TextSTAT, which is very user-friendly and its use and application is free of charge. It is a simple programme for the analysis of texts. It reads plain text files (in different encodings) and HTML files (directly from the internet) and it produces word frequency lists and concordances from these files.
(Fig. 2).

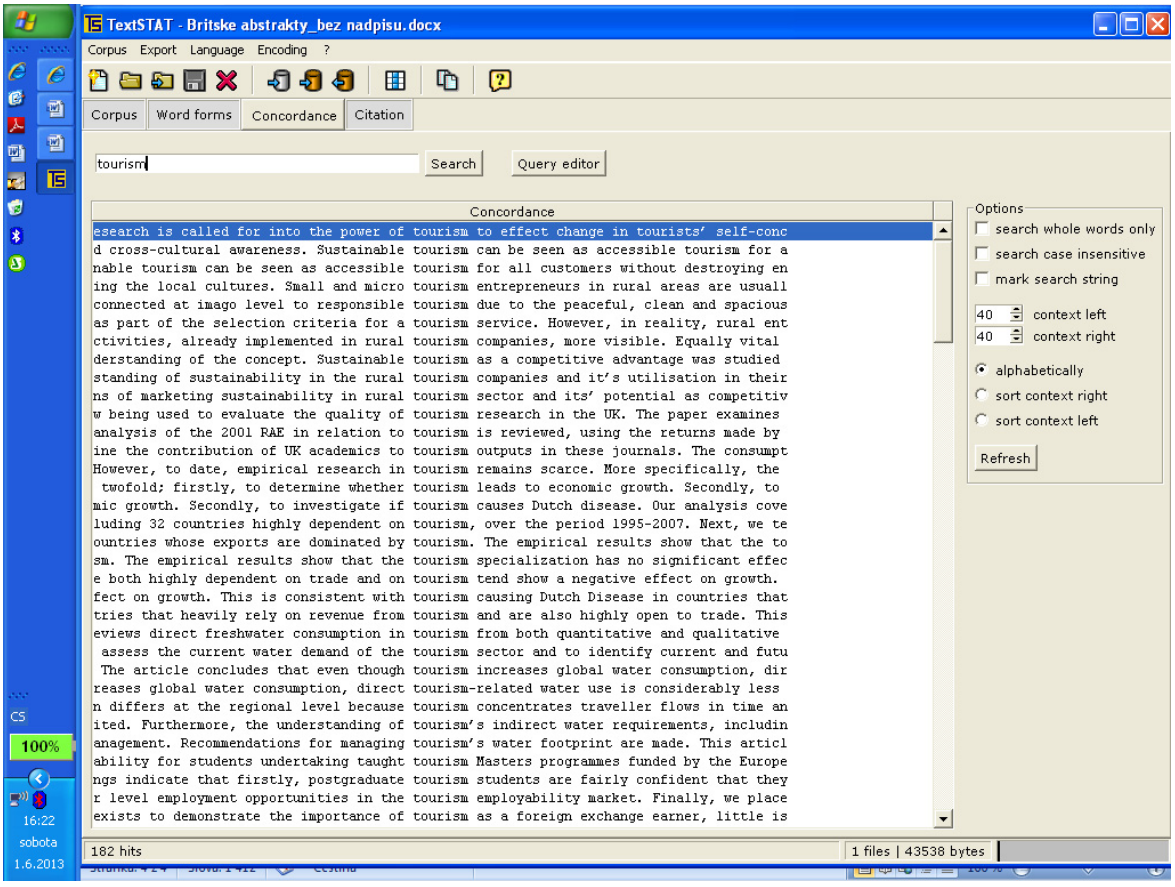

Fig. 2. TextSTAT concordances of the word tourism

Thus, on the basis of the above stated theory, to enhance the development of writing skills, teachers can do the following activities with their students:

- students can see real texts of everyday use, such as letters, essays or reports;
- thanks to the WordList, they can see which words are most common in the writing, e.g., of formal letters;
- thanks to the concordance, they can explore the most common collocations of the word tourism (see above).

In conclusion, corpus linguistics and particularly, the software programmes used for the analysis of different corpora represent an invaluable tool for foreign language teachers.

References

Greenbaum, S. (1991).The development of the international corpus of English. London: Longman.
Scott, M. (2012). WordSmith Tools. Version 6.0. UK.

TextSTAT. (2012). Retrieved March 11, 2013, from http://neon.niederlandistik.fu-berlin.de/en/textstat/.
Wu. M. H. (1992). Towards a contextual lexico-grammar: an application of concordance analysis in EST teaching. *RELC Journal,* 23(2), 18-34.